A Comparative Study in Corpus Linguistics applied to Automatic Terminology Extraction

Mercè Vàzquez, Sergi Alvarez-Vidal, Antoni Oliver

Universitat Oberta de Catalunya, Universitat Autònoma de Barcelona, Universitat Oberta de Catalunya
Barcelona (Spain)
mvazquezga@uoc.edu, sergi.alvarez@uab.cat, aoliverg@uoc.edu

Abstract

Parallel and comparable corpora are the main linguistic resources to identify multilingual terminology using automatic term extraction tools. However, parallel corpora are available only for certain languages, domains and genres, and comparable corpora have some limitations when identifying corresponding terms. To implement a more effective selection of multilingual terminology, we compared the performance of using specialised parallel and comparable corpora applied to languages with various forms of capital in linguistic resources. This paper presents a comparative study in corpus linguistics in which we automatically identify terms in Catalan, Spanish and English in Legislation and Administrative Law using parallel corpora, comparable corpora and a combined methodology based on both types of corpora together with word embeddings. We observe that the combined methodology implemented obtains a higher number of terms than when working exclusively with parallel or comparable corpora. The evaluation of the results is performed using a terminological thesaurus as a gold standard. The new methodology presented in our study permits us to identify multilingual terminology in an effective way, especially in Catalan-Spanish languages.

Keywords: comparable corpora, parallel corpora, automatic terminology extraction, computational terminology

1. Introduction

Terms are elements within specialised documents that are used for the creation and enrichment of ontologies and dictionaries (Maynard & Ananiadou 2001; El-Sappagh et al. 2018; Durán-Muñoz, 2019), and terms are relevant for multitude of applications such as information retrieval (Lingpeng et al., 2005), machine translation (Haque et al., 2019; Michon et al., 2020; Moslem et al., 2023), and sentiment analysis (Mayorov et al., 2015). Since the nineties computational terminology has been widely developed through the availability of corpus linguistics and the contribution of different natural language processing (NLP) domains, such as terminology extraction, information retrieval, ontology building, machine translation or computer-aided translation (L'Homme et al., 1998).

In order to implement different ATE strategies to identify terms from specialised information, monolingual and parallel corpora - a corpus that contains source texts and their translations (McEnery & Xiao, 2007) - have become the main linguistic resources to automatically extract candidates to compile terminology from a specific domain to be manually supervised by linguists and terminologists (Kupiec, 1993; Gaussier, 1998; Ha et al., 2008; Macken et al., 2013; Haque et al., 2014; Baisa et al., 2015). However, parallel corpora are usually available only for certain languages, domains and genres, and term extraction tools have been developed exclusively for the needs of major European and non-European languages. Furthermore, compiling parallel corpora from authoritative sources of information for terminology extraction is a resource-intensive task, particularly for less-resourced languages in terms of data scarcity in specialised domains. Indeed, access to authoritative sources sometimes may be restricted, or may need the permission from authors, companies or institutions (Daille & Morin, 2005; Gornostay et al. 2012; Gurrutxaga et al. 2013; Rigouts Terryn et al., 2018). Consequently, comparable corpora – documents that are comparable in content and form in various degrees and dimensions across several languages or language varieties (Zweigenbaum, Rapp and Sharoff, 2024) – are an alternative for extracting domain-specific terms because it is much easier to collect data (Fung, &Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Daille & Morin, 2005; Aker et al., 2013; Bouamor et al., 2013; Morin & Hazem, 2014; Hazem & Morin, 2016; Hazem & Morin, 2017). However, certain limitations have been observed when implementing automatic term extraction from comparable corpora related to the terms identification, as the information compiled in both languages is similar but not equivalent, due to comparable corpora are not aligned; the construction of a gold standard dataset to automate the task, and also the evaluation of candidates term extracted using comparable corpora (Rigouts Terryn et al., 2020; Rigouts Terryn, 2023).

In order to make a more effective selection of terminology from corpus linguistics, we analysed the performance of specialised parallel and comparable corpora applied to languages with various forms of capital in linguistic resources**.** We present a comparative analysis on corpus linguistics in which we automatically identify terms in Catalan, Spanish and English in Legislation and

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 20–29
11 May 2026.

Administrative Law using parallel corpora, comparable corpora and a combined methodology based on both types of corpora together with word embeddings. This comparison aims to determine whether a particular corpora type is more suitable for processing using ATE tools. To conduct this analysis, we utilise TBXTools, an open-access term extraction tool capable of employing both statistical and linguistic term extraction methods and automatic search of translation equivalents of terms in corpora (Oliver and Vàzquez, 2015), to extract candidate terms from comparable and parallel corpora in the Legislation and Administrative Law in order to assess the reliability of the results obtained.

The primary objective of our study is to assess the performance and reliability of comparable corpora in comparison with parallel corpora to automatically identify multilingual terminology using a term extraction tool, particularly in languages with limited linguistic resources. This primary objective is based on two hypotheses. The first is that comparable corpora allows us to easily compile a higher volume of specialised information compared to parallel corpora, due to the fact that collecting original texts in more than one language for one domain is easier to compile than a collection of translated and aligned texts. The second is that comparable corpora can be an effective and reliable mechanism to compile specialised information for all domains, genres and languages, which is especially important for less-resourced languages.

To achieve this aim, this paper conducts a comparative analysis of results obtained using parallel corpora from the same domain to ascertain the effectiveness and reliability of Legislation and Administrative Law comparable corpora. The findings will enable us to determine whether employing comparable corpora (a) yields a larger corpus volume for terminology extraction, (b) improves terminology identification in languages with restricted linguistic resources such as Catalan, and (c) achieves a satisfactory level of terminological reliability.

The remainder of the present paper is structured as follows: in Section 2 we describe the background of parallel and comparable corpora applied to terminology extraction. In Section 3 we present the materials and tools used and the method implemented to compare the performance of comparable corpora with parallel corpora to identify multilingual terminology. The results and discussion are described in detail in Section 4. The paper is concluded with some final remarks and ideas for future research.

2. Materials, tools, and methods

With the aim of making a more effective selection of term equivalents from corpora using ATE tools, we analyse and compare the performance of specialised parallel and comparable corpora applied to automatic terminology extraction for two language pairs (English-Spanish and Catalan-Spanish) in one domain, Legislation and Administrative Law.

2.1 Materials

We have processed two parallel corpora to obtain parallel and comparable corpora of different sizes. For the English-Spanish language pair, we have used the DGT Corpus (Steinberger, 2013), and for the Catalan-Spanish, the DOGC corpus, a Catalan Spanish parallel corpus created from laws of the Catalan government (Oliver, 2022). The *Diari Oficial de la Generalitat de Catalunya*[1] (DOGC) is an official media outlet in which the laws and regulations of the Government of Catalonia are published. In Table 1 we can observe the size of these corpora once the segments have been deduplicated and shuffled. These corpora contain unique sentences.

Corpus	Segments	L1 tokens	L2 tokens
DGT unique eng-spa	3,640,761	73,883,784	84,702,886
DOGC unique cat-spa	8,472,786	188,929,206	197,986,300

Table 1: Tokens and segments included in the used corpora

From these corpora we have created one subset of parallel corpora: 1M segments. To create the comparable corpora from the parallel corpora we have selected 1M segments from the top of the corpus for the source language and from the bottom of the corpus for the target language. We have then a comparable corpus of 1M segments for each language pair.

For our study, we also need a set of source terms to find their translation, and the valid translation equivalents to perform the evaluation of the methodologies. We have used a subset of the Catalan IATE terminology glossary[2] consisting of 1,722 terms in English, Spanish and Catalan extracted from the Europarl Corpus (Koehn, 2005). More precisely, the glossary has 1,621 terms with equivalents in English and Spanish; and 1,232 with equivalents in Catalan and Spanish.

We have created subsets of this glossary with the terms present in all the created parallel and comparable corpora to evaluate the performance of terminology extraction using parallel corpora,

[1] https://dogc.gencat.cat/ca/inici/

[2] https://www.termcat.cat/en/diccionaris-en-linia/264

comparable corpora and also a combined methodology. From each source and target terms we also know the frequency of apparition of each source and target term. This data will help us to provide a more detailed analysis of the evaluation figures. In Table 2 we can see the number of source-target terms present in each subcorpus.

Language	Corpus	Size	Terms
eng-spa	parallel	1M	617
eng-spa	comparable	1M	852
cat-spa	parallel	1M	814
cat-spa	comparable	1M	845

Table 2: Number of terms present in the parallel and comparable corpora used in the experiments

2.2 Tools

To undertake the comparative analysis, we have used TBXTools, a Python class performing a series of methods for automatic term extraction and automatic search of translation equivalents of terms in parallel and comparable corpora. For the experimental part we have used the capabilities of TBXTools for automatic detection of translation equivalents of known terms in parallel and comparable corpora.

2.2.1 Detection of translation equivalents in parallel corpora

In TBXTools the automatic detection of translation equivalents in parallel corpora is performed in the following way: we have a set of terms in the source language (L1) and we want to know the translation equivalents of these terms in the target language (L2). We have a parallel corpus L1-L2 for the given subject. The algorithm takes one term in L1 and creates a L2 subcorpus with the target segments whose source segments contain the given term. Then the algorithm performs an ATE task (it can be either statistical or linguistic) on this L2 subcorpus. The most frequent L2 term candidate has big chances of being the translation equivalent of the given L1 term. We repeat this procedure for each term in the set of terms in L1.

In this strategy, the ATE process on the target corpus can be either statistical or linguistic. In our experiments, the statistical methodology has been used. One important parameter is the relation of n-gram order (*n*) between the source term and the target term candidate. For example, one uni-gram source term can have a uni-gram translation equivalent (as in *agreement - acuerdo*), a bi-gram (as in *bailiff - agente judicial*) and even a tri-gram (as in *affidavit - acta de notoriedad*). As these relations are not known in advance, several relations should be explored. Therefore, two parameters: maximum *n* increment (max_inc) and maximum *n* decrement (max_dec) should be set to better identify the translation equivalents.

This simple strategy works quite well when the L1 terms appear several times in the L1 part of the parallel corpus and most of the times the same translation equivalent is used in the L2 part of the parallel corpus. In the experimental part we present figures of precision, recall and F_1 for this strategy in several scenarios.

2.2.2 Detection of translation equivalents in comparable corpora

In TBXTools a method for automatic detection of translation equivalents in comparable corpora based on word embeddings, is implemented. In this methodology the embeddings for all terms in the list of terms to search is calculated using the source language part of the comparable corpus. As the source language terms are known, we can convert the complex terms, that is, terms formed by more than one word, into single tokens joining the components of the terms by some symbol, for example "_". In this way, a complex term as for example *interest rate* is converted into a single token *interest_rate*. We call this process *compoundifying*.

Hence, source language terms can be compoundified because they are known, as they are in the list of terms we want to find the translation equivalent from. Then, we need to calculate the embedding for the target terms using the comparable corpus for the target language. But now these target language terms are not known in advance, as we are precisely looking for these terms. This is not a problem for the simple terms, that is, for the terms formed by a single word, as we can calculate the embeddings for all the words in the target comparable corpus. But we don't know *a priori* the complex terms in the target language, so the algorithm performs an ATE process in the target comparable corpora to detect target term candidates in order to compoundify them. This ATE process, again, can be either statistical or linguistic. We implemented the statistical process.

Once we have the embeddings for the source terms and the target extracted terms, we have two different vector spaces that should be mapped. To do so, we use the vecmap[3] algorithm (Artetxe et al., 2018). Once the two vector spaces are mapped, the translation equivalent of a source language term should be the nearest target language term in the mapped vector space. But taking the nearest target language term is dangerous, as this target term can be closer to a different source term. For this reason, a margin score is calculated, as defined in Artetxe and Schwenk (2019) (changing the sentence embeddings by word embeddings): the margin score between two candidate term equivalents *x* and *y* is defined as the ratio between the cosine distance between the two word embeddings, and the average cosine similarity of its nearest neighbors in both directions. This strategy, however, fails in the cases where the translation

[3] https://github.com/artetxem/vecmap

equivalent is not present in the target comparable corpus. For this reason, a minimum margin score should be defined to reject the equivalents detected with a lower margin score.

2.3 Methods

The methods we implemented to identify a more effective procedure to select term equivalents from corpora are based on parallel corpora, comparable corpora and a combined methodology which combines both types of corpora.

2.3.1 Parallel corpora

In this strategy, as stated above, two important parameters must be set: max_dec and max_inc. When searching for the translation equivalent of a term, the translation equivalent might have the same number of tokens, or a different number. For example, the translation of a bigram term can be a unigram (max_dec=1) or a trigram (max_inc=1). Depending on the language pair in the experiments, these parameters can be set with different values. To set these parameters in our experimental setting we have performed a statistical analysis using the complete Catalan IATE e-dictionary (Vàzquez, Oliver, 2019) from Termcat's Terminologia Oberta. This e-dictionary contains 15,997 terms in Catalan, Spanish, English and French. From this e-dictionary we have analyzed all the English-Spanish and Catalan-Spanish pairs of terms to count for the increments and decrements in the n-gram relation between source and target term. In Table 3 we can observe the results of this analysis, and we can set for English-Spanish max_dec=1 and max_inc=1 and for Catalan-Spanish we can set max_dec=0 and max_inc=0.

Increment	% English-Spanish	% Catalan-Spanish
-3	2,16	2.95
-2	4.75	6.21
-1	10.46	4.32
0	60.04	69.55
1	12.88	5.21
2	4.04	3.2
3	1.1	1.1

Table 3: Results of the statistical analysis to set the max_dec and max_inc parameters

For each source term the algorithm provides a set of translation equivalents sorted by a confidence score, being the first candidate the one with more chances to be the correct one.

2.3.2 Comparable corpora

To find the translation equivalents of terms in comparable corpora, we need to perform the following processes:

1. Compoundify the source language terms in the source language comparable corpus. This process can be performed as the source language terms are known in advance to evaluate the performance of terminology extraction using comparable corpora.
2. Compoundify the target language terms in the target language comparable corpus. As the target language terms are not known in advance, we should make an unsupervised ATE process in the target corpus to get a set of term candidates to compoundify.
3. Calculate the source language word embeddings using the compoundified source language comparable corpus.
4. Calculate the target language word embeddings using the compoundified target language comparable corpus
5. Map the source and target language embeddings.
6. Extract the target term candidate for each source language term, using the margin score.

2.3.3 Combined methodology

In the combined methodology the translation equivalent candidates are obtained using the parallel corpora method, but the translation equivalents will be sorted using mapped word embeddings calculated by concatenating the source part of the parallel corpus with the comparable corpus for the source language, and the target part of the parallel corpus with the comparable corpus for the target language. The steps performed in this methodology are the following:

1. Perform the search of translation equivalents using the parallel corpus method.
2. Concatenate the source part of the parallel corpus and the comparable corpus for the source language.
3. Concatenate the target part of the parallel corpus and the comparable corpus for the target language.
4. Compoundify the concatenated corpus for the source language using the source terms we want to search for.
5. Compoundify the concatenated corpus for the target language using all the translation equivalents candidates obtained in the first step.
6. Calculate the source language embeddings using the compoundified source language concatenated corpus.
7. Calculate the target language embeddings using the compoundified target language concatenated corpus.
8. Map the source and target language embeddings.
9. Resort the translation equivalents calculated in step one calculating the margin score.

The combined methodology has the additional advantage that the target language corpora can be compoundified without the need of performing an unsupervised ATE process. Instead, we can compoundify the target language corpora using all the translation equivalents candidates obtained using the parallel corpus methods, as these candidates are precisely the ones we want to resort with the margin score.

3. Results

In this section we present the evaluation results for three tasks: automatic translation equivalent detection in parallel corpora, in comparable corpora and a combined methodology using parallel corpora and embeddings calculated from the parallel corpus and a comparable corpus to resort the candidates. These methodologies are tested for 1M segments of the corpora and two language pairs: English-Spanish and Catalan-Spanish. We show the results obtained with 1M corpus segments.

3.1 Parallel corpora

3.1.1 English-Spanish

The evaluation figure for the automatic detection of translation equivalents in parallel corpora for English-Spanish and corpus size of 1M segments are shown in Table 4. For this corpus size we experimented with two sets of max_dec=0 and max_inc=0, and max_dec=1 and max_inc= 1, which offers higher results. In the Table we can observe the precision (P), recall (R) and F_1 for the first translation candidate, and for the cases where the correct candidate is among the first 5 candidates (P 5, R 5 and F_1 5), and among the top-ten candidates (P 10, R 10 and F_1 10). We also present figures for those source terms appearing at least 1, 2, 5, and 10 times in the corpus. As a general behavior, the most frequent the source term is, the higher the precision. But as the recall has been calculated regardless the frequency of apparition, the recall and F_1 score drop drastically.

Another general and obvious behavior is that the P5 and P10 (the precision taking into account the top 5 or 10 candidates) is higher than P1 (the precision taking into account only the first candidate). But it is worth knowing P5 and P10, because it simulates the practical case where a terminologist is presented with the list of candidates to choose the correct one.

An interesting conclusion from Table 4 is that enlarging the size of the parallel corpora does not improve the results. Enlarging the size of the parallel corpora causes some source terms to appear with a higher frequency, but some new source terms with lower frequency are also included in the experiment, yielding no improvement.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	21.88	21.88	21.88	42.46	42.46	42.46	50.57	50.57	50.57
2	26.08	21.56	23.6	50.39	41.65	45.61	58.43	48.3	52.88
5	32.44	19.61	24.44	57.64	34.85	43.43	64.61	39.06	48.69
10	36.77	17.34	23.57	62.89	29.66	40.31	70.1	33.06	44.93

Table 4. Evaluation figures for parallel corpus 1M segments English-Spanish with max_dec=1 and max_inc=1

3.1.2 Catalan-Spanish

The results for the automatic detection of translation equivalents in parallel corpora for Catalan-Spanish in 1M segments corpora are presented in Table 5. For this language pair we have only considered max_dec and max_inc of 0. The first thing we notice is that the results for Catalan-Spanish are much better than for English-Spanish. This can be explained by different causes. The 0 value of max_inc and max_enc covers a larger percentage of cases for Catalan-Spanish than for English-Spanish. The fact that Catalan-Spanish are more similar than English-Spanish should have no direct influence on the results, as no linguistic information is used. But the similarity between languages may cause a more consistent use of translation equivalents in the corpus, making them easier to detect.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	44.72	44.72	44.72	78.26	78.26	78.26	83.78	83.78	83.78
2	48.25	44.1	46.08	83.74	76.54	79.97	88.04	80.47	84.08
5	51.89	40.54	45.52	88.05	68.8	77.24	91.19	71.25	80.0
10	53.61	36.49	43.42	88.81	60.44	71.93	91.88	62.53	74.42

Table 5. Evaluation figures for parallel corpus 1M segments Catalan-Spanish with max_dec=0 and max_inc=0

3.2 Comparable corpora

3.2.1 English-Spanish

The evaluation results for comparable corpora for the English-Spanish pair are now presented. As explained in section 2.3.2, one important step in this methodology is the compoundifying of complex terms, that is, converting the terms formed by more than one word, into a single token, replacing the blank spaces by a "_". As commented, this process can be done for the source terms, as they are already known. But target terms are still not known, so we cannot directly compoundify them.

We present two cases:

1. False compoundifying, where we cheat and use the list of known target terms used for evaluation. The results obtained are higher than in a real situation, but allow us to assess the capability of mapped word embeddings to find translation equivalents.
2. Statistical compoundifying, where an unsupervised statistical term extraction has been performed on the target comparable corpus. We then take all the target term candidates and we use them to compoundify. In Table 6 we can observe the overall number of term candidates and the number of terms with a frequency of 5 or higher, used in the compoundifying process.

	Term candidates	
Corpus size	freq. >=1	freq. >=5
1 M	943,544	238,005

Table 6. Number of term candidates of the automatic term extraction for compoundifying the Spanish comparable corpora in the English-Spanish experiments

False compoundifying

In Table 7 we can observe the evaluation results for the English-Spanish experiments using comparable corpora, but doing the cheating of compoundifying the target Spanish terms, for the size of 1M segments. The results for 1M segment corpus are low, but with about 10 points of increment in precision for the first candidate and for the first 5 candidates; and up to 20 points for the first 10 candidates. Still, however, the precision results are not enough for fully automatic tasks, but they can be reliable enough for manual tasks performed by terminologists to provide a set of suggestions.

But we must remember that the results from Table 7 are obtained using the known Spanish translation candidates to compoundify the target corpus. In most real situations this is not doable, as these Spanish translation equivalents are still unknown. Only the case when a terminologist performs an ATE task and a manual revision of the term candidates, and then tries to search the relation with the set of source English terms would be similar to this experimental setting.

Freq.	P	R	F_1	P 5	R 5	F_1 5	F 10	R 10	F_1 10
1	13.02	12.91	12.96	18.46	18.31	18.39	40.24	39.91	40.07
2	13.02	12.91	12.96	18.46	18.31	18.39	40.24	39.91	40.07
5	15.44	15.33	15.39	21.95	21.8	21.88	47.73	47.4	45.57
10	17.38	17.27	13.33	24.24	24.09	24.17	52.79	52.46	52.62

Table 7. Evaluation results for the comparable 1M corpora using false compoundifying

Compoundifying with statistical ATE

In Table 8 the results of the search of translation equivalents using comparable corpora, and compoundifying using an unsupervised statistical ATE task on the target Spanish comparable corpus are presented. This experimental setting is more similar to a real practical situation. These results cannot be used in a full automatic setting, but can be presented as an aid to a terminologist.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	4.97	4.93	4.95	9.59	9.51	9.55	24.38	24.18	24.28
2	4.97	4.93	4.95	9.59	9.51	9.55	24.38	24.18	24.28
5	5.95	5.91	5.93	11.47	11.39	11.43	29.18	28.97	29.08
10	6.7	6.66	6.68	12.92	12.84	12.88	32.54	32.33	32.43

Table 8. Evaluation results for the comparable 1M comparable corpora using unsupervised statistical ATE for compoundifying

3.2.2 Catalan-Spanish

In this section we present the evaluation of the task on comparable corpora for Catalan-Spanish. As we have done for English-Spanish, we present the results with two compoundifying processes: firstly, doing the cheating of using the already known Spanish terms; and secondly, performing an unsupervised statistical ATE process on the Spanish comparable corpora, and using the term candidates with a frequency equal or higher than 5 to compoundify the target corpus. In Table 9 we can see the number of extracted terms and those used for compoundifying.

	Term candidates	
Corpus size	freq. >=1	freq. >=5
1 M	759,116	209,872

Table 9. Number of term candidates of the automatic term extraction for compoundifying the Spanish comparable corpora in the English-Spanish experiments

False compoundifying

In this setting, the results shown in Table 10 are much better than the results for English-Spanish (see Table 7). For the precision of the first candidate we obtain an improvement of 21.23 for the 1M segments corpora. For Catalan-Spanish we get precisions of around 90% with very good F_1 scores if we take into account the top-ten candidates for source terms appearing at least 10 times. But we must keep in mind that the compoundifying step has been performed with the cheating of using the Spanish terms in the evaluations set, and these good results will not be obtained in a real situation.

Compoundifying with statistical ATE

Now, in a more realistic situation where the compoundifying step has been performed through unsupervised statistical ATE (Table 11), the precision values drop drastically, but they are much better than the obtained for the English-Spanish corpora (see Table 8), with an improvement of 16.92 precision points for the 1M segments corpora.

3.3 Combined method using word embeddings

3.3.1 English-Spanish

From the results presented so far, we can see that the precision values obtained using the methodology based on parallel corpora is much higher than those based on comparable corpora. The main problems for the methodology based on parallel corpora are, on one hand, to determine the translation equivalent for terms with very low frequency; and on the other hand, to know the n-gram relation between the source term and the target term. In this section we explore the use of mapped word embeddings to resort the list of translation equivalents. In this combined methodology we use both the parallel and the comparable corpora to calculate the word embeddings. Then, for each source term, we get the list of translation equivalent candidates using the parallel corpus. Once we get the list of the *n* best candidates, we resort them using the margin score calculated with the source and target word embeddings.

In Table 12 we can observe the results using the parallel and comparable corpora with 1M segments. These results should be compared with the results presented in Table 4. Note that we are using the worst values of max_dec and max_inc parameters, as the embeddings will do the job of selecting the correct translation equivalent regardless of the *n* order relation. Comparing these two tables we can see that this reordering methodology using word embeddings is very productive for the first candidates. If we analyze the results taking into account the precision of the first candidate (P), we get an improvement of 19.1 points for a frequency of 1 (terms appearing at least one time), an improvement of 16.14 points for frequency of 5, and 15.13 for a frequency of 10. These figures drop when considering the first 5 candidates (P5) to 9.83, 4.71 and 3.78 respectively. But when we observe the results for the top-ten candidates (10), we improve a little (3.32 points), but get worse results for frequency 5 (-1.86 points) and frequency 10 (-3.87 points). This is explainable because resorting a large set of candidates has no effect if we take all of them to calculate the precision. So the results seem to indicate that the methodology can be very productive to select the best candidate into the first position when considering a limited number of candidates.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	34.25	32.43	33.31	37.38	35.38	36.35	75.75	71.72	73.68
2	34.24	32.43	33.31	37.38	35.38	36.35	75.75	71.72	73.68
5	38.32	36.53	37.41	41.68	39.73	40.68	84.48	80.83	82.46
10	40.94	38.99	39.94	43.84	71.74	42.76	88.58	84.35	86.41

Table 10. Evaluation results for the comparable 1M corpora using false compoundifying for Catalan-Spanish

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	21.89	19.76	20.77	27.92	25.21	26.49	59.24	53.49	56.22
2	21.89	19.76	20.77	27.92	25.21	26.49	59.24	53.49	56.22
5	24.13	22.13	23.09	30.81	28.27	29.49	65.12	59.73	62.31
10	26.06	24.06	25.02	33.12	30.58	31.8	69.54	64.2	66.77

Table 11. Evaluation results for the comparable 1M comparable corpora using unsupervised statistical ATE for compoundifying

Freq.	P	R	F_1	P 5	R 5	F_1 5	P10	R 10	F_1 10
1	40.98	21.72	28.39	52.29	27.71	36.23	53.82	28.53	37.29
2	44.0	21.39	28.79	56.0	27.23	36.64	57.67	28.04	37.73
5	48.58	19.45	27.78	62.35	24.96	35.65	62.75	25.12	35.88
10	51.9	17.67	23.36	66.67	22.69	33.86	67.14	22.85	34.1

Table 12. Evaluation figures for parallel corpus 1M segments English-Spanish with max_dec=1 and max_inc=1 combined methodology

3.3.2 Catalan-Spanish

This combined methodology is also very productive for the Catalan-Spanish pair. If we observe the results in Table 13 for the 1M segments corpora and compare the results in Table 5 for the parallel corpus methodology for the same corpora size, we see an improvement of 31.85 precision points for the first candidate of

frequency 1 or higher. As in this combined methodology the results of the parallel corpus methodology are resorted with the mapped word embeddings calculated with both the parallel and comparable corpora, the improvements are much higher for the first position, dropping to 4,06 points for the first five candidates and yielding to no improvement for the top-ten candidates.

Freq.	P	R	F_1	P 5	R 5	F_1 5	P 10	R 10	F_1 10
1	76.57	71.87	74.14	82.33	77.27	79.72	83.77	78.62	81.12
2	81.43	70.02	75.3	87.14	79.94	80.58	88.43	76.04	81.77
5	85.26	63.27	72.64	90.23	66.95	76.87	91.06	67.57	77.57
10	86.34	55.9	67.86	91.08	58.97	71.59	91.65	59.34	72.04

Table 13. Evaluation figures for parallel corpus 1M segments Catalan-Spanish with max_dec=0 and max_inc=0 combined methodology

If we now compare the improvements of the combined methodology for Catalan-Spanish and English-Spanish (comparing Table 12 with Table 13), we can observe higher improvements for Catalan-Spanish (31.85 precision points vs. 19.1 points for the first position and frequency 1).

4. Conclusions and perspectives

In this article, we have presented a novel usage of parallel and comparable corpora to effectively identify multilingual terminology from specialised domains. The methodology combines the information contained in parallel corpora and comparable corpora related to a specific domain and introduces a mapped word embeddings procedure to effectively identify term equivalents from specialised corpora. In order to determine the reliability of the method, especially addressed to less-resourced languages that suffer from a lack of available linguistic resources to build parallel corpora, we have conducted a comparative analysis on corpus linguistics in which we automatically identify terminology in Catalan, Spanish and English in Legislation and Administrative Law using parallel corpora, comparable corpora and a combined methodology based on both types of corpora together with word embeddings. The evaluation results applied in two language pairs (English-Spanish and Catalan-Spanish) in the domain of Legislation and Administrative Law shows that combining parallel and comparable corpora to identify terminology from specialised domains outperforms those using parallel corpora or comparable corpora separately. We have used a terminological glossary manually compiled as a gold standard from the Catalan IATE e-dictionary to evaluate the reliability of the method applied. The promising results obtained contribute to expanding the methodology applied in corpus linguistics to maximise the terminology compilation, which has a relevant impact in the context of less-resourced languages with a lack of corpus linguistics availability.

The novel usage of corpus linguistics has been implemented in TBXTools, an open-access term extraction tool created to automatically identify multilingual terminology from specialised domains. The present methodology can be used in any other specialised area that has similar resources to identify terminology.

The present research provide a promising perspective in terminology identification with a novel usage of corpus linguistics with the aim to provide a larger volume of corpora for terminology extraction, especially relevant in the context of languages with limited linguistic resources; enhance terminology detection, and achieve a satisfactory level of reliability in the extracted terminology.

As a future work, we plan to introduce general-domain data to improve translation terms identification, due to general content completing the information given for each term candidates in source and target corpora. And also we plan to evaluate the performance of the methodology applied in other domains together with other evaluation methods.

5. Acknowledgments

This work is supported by the project TamTAS PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and European Union in the Chist-era call 2025 Science in your own language

6. Bibliographical References

Aker, A., Paramita, M. and Gaizauskas. R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp 402–411). Sofia, Bulgaria. Association for Computational Linguistics.

Artetxe, M. Labaka, G. and Agirre. E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp 789–798).

Artetxe, M., and Schwenk. H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, *7*, 597–610.

Baisa, V., Ulipová, B. and Cukr. M. (2015). Bilingual terminology extraction in Sketch Engine. *9th Workshop on Recent Advances in Slavonic Natural Language Processing*, 61–67.

Bouamor, D., Semmar, N. and Zweigenbaum. P. (2013). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp 759–764).

Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Chiao, Y. C., and Zweigenbaum. P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics* (pp 1–5). Association for Computational Linguistics.

Daille, B., and Morin. E. (2005). French-English terminology extraction from comparable corpora" *International Conference on Natural Language Processing*, 707–718. Berlin, Heidelberg: Springer Berlin Heidelberg.

Durán-Muñoz, I. (2019). Methodological Proposal to Build a Corpus-Based Ontology in Terminology" *Lingue e Linguaggi*, *29*, 581–597.

El-Sappagh, S., Franda, F., Ali, F., and Kwak K. S. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC medical informatics and decision making*, *18*, 1–19.

Fung, P., and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp 414–420).

Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp 444–450).

Gornostay, T., Ramm, A., Heid, U., Morin, E., Harastani R., and Planas E. (2012). Terminology Extraction from Comparable Corpora for Latvian. *HLT 2012: 5th International Conference Human Language Technologies*, 66–73. Estonia.

Gurrutxaga, A., Leturia, I., Saralegi, X., and Vicente, I. S. (2013). Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making. *Building and using comparable corpora*, 51–75.

Ha, L. A., Fern, G., Mitkov, R., and Corpas, G. (2008). Mutual bilingual terminology extraction" In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08) (pp 1818–1824).

Haque, R., Penkale, S., and Way, A. (2014). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology* (Computerm) (pp 42–51).

Haque, R, Hasanuzzaman, Md., and Way, A. (2019). Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP 2019) (pp 437–446). Varna, Bulgaria.

Hazem, A., and Morin, E. (2016). Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. *26th International Conference on Computational Linguistics* (COLING), 3401–3411. Osaka, Japan.

Hazem, A., and Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (pp 685–693).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *The 10th Machine Translation Summit Proceedings of Conference*. 79–86. International Association for Machine Translation.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* (pp 17–22). Association for Computational Linguistics.

L'Homme, M.-C., Bourigault, D., and Jacquemin, C. (1998). *First Workshop on Computational Terminology* (COMPUTERM'98), Montréal, Canada.

Lingpeng, Y., Donghong, J., Guodong, Z., and Yu, N. (2005). Improving Retrieval Effectiveness by Using Key Terms in Top Retrieved Documents. In Losada, D.E., Fernández-Luna, J.M. (Eds.) Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, *3408* (pp 169–184). Springer, Berlin, Heidelberg.

Macken, L., Lefever, E., and Hoste, V. (2013). ExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-Based Alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *19*(1), 1–30.

Maynard, D., and Ananiadou, S. (2001). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing*, *8*(1), 101–125.

Mayorov, V., Andrianov, I., Astrakhantsev, N., Avanesov, V., Kozlov, I., and Turdakov. D. (2015). A High Precision Method for Aspect Extraction in Russian. *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue."*, *2*, 34–43. Moscow, Russia.

McEnery, A., and Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to. *Incorporating corpora: Translation and the linguist. Translating Europe. Multilingual matters*, 1–13.

Michon, E., Crego, J., and Senellart, J. (2020). Integrating Domain Terminology into Neural Machine Translation. In P*roceedings of the 28th International Conference on Computational Linguistics* (pp 3925–3937). Barcelona, Spain. International Committee on Computational Linguistics.

Morin, E., and Hazem, A. (2014). Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp 1284–1293).

Moslem, Y., Romani, G., Molaei, M., Kelleher, J. D., Haque, R., and Way, A. (2023). Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In *Proceedings of the 8th Conference on Machine Translation* (pp 902–911). Singapore. Association for Computational Linguistics.

Oliver, A. 2022. El corpus paral·lel del Diari Oficial de la Generalitat de Catalunya. *Linguamática*, 2023, 14 (2).

Oliver, A., and Vàzquez. M. (2015). TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp 473–479).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp 519–526).

Rigouts Terryn, A., Hoste, V., and Lefever, E. (2020). In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, *54*(2), 385–418.

Rigouts Terryn, A. (2023). Supervised Feature-based Classification Approach to Bilingual Lexicon Induction from Specialised Comparable Corpora. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)* (pp. 59–68).

Steinberger, R., Eisele A., Klocek, S., Pilos, S. and Schlüter, P. (2013). *DGT-TM: A freely available translation memory in 22 languages. European Commission*.

Vázquez, M., Oliver, A., and Casademont, E. (2019). Using open data to create the Catalan IATE e-dictionary. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 25*(2) (pp. 175–197).

Zweigenbaum, P., Sharoff, S., & Rapp, R. (2024). Preface. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora*. LREC-COLING-2024.